Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Department of Mathematics and Physics

Master of Data Science

Multi-Dimensional Clustering of Roles in the NBA

THESIS to obtain the **DEGREE** of
MASTER OF DATA SCIENCE

A thesis presented by:
Elijah Daniel Stutzman

Thesis Advisor:
Dra. Rocío Carrasco Navarro

Tlaquepaque, Jalisco, May, 2021

Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Department of Mathematics and Physics

Master of Data Science Approval Form

Thesis Title: **Multi-Dimensional Clustering of Roles in the NBA**

Author: **Elijah Daniel Stutzman**

Thesis Approved to complete all degree requirements for the Master of Science Degree in Data Science.

Thesis Advisor, **Dra. Rocío Carrasco Navarro**

Thesis Reader, **Dr. Luis Fernando Luque Vega**

Thesis Reader, **Dr. Esteban Jiménez Rodríguez**

Academic Advisor, **Ms. Juan Carlos Martínez Alvarado**

Tlaquepaque, Jalisco, May, 2021

Multi-Dimensional Clustering of Roles in the NBA

Elijah Daniel Stutzman

Abstract

While in the National Basketball Association (NBA), players are often described by the position that they play and not necessarily the role that they fill on the team. In this thesis, newly defined player roles have been identified by applying multi-dimensional clustering techniques on thirty-eight variables for over ten thousand player samples. These roles help to differentiate players that play the same traditional position, and will allow for new comparisons between players to be produced. Using player statistics from nineteen seasons, models were developed using three separate clustering techniques: Gaussian Mixtures, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and k-Means. After the models were developed a final model was chosen that provided the best clusters that were used to identify the new roles. These new roles are able to be used to identify replacements for certain players, signing a player that fulfills the same role, or by drawing comparisons between new players in the NBA and the historical roles that other players have fulfilled.

Multi-Dimensional Clustering of Roles in the NBA

Elijah Daniel Stutzman

Abstract

En la NBA, es comun que los jugadores se identifican por la posición que juegue, no necesariamente el rol que toman en el equipo. En esta tesis, se define nuevos roles de jugadores aplicando tecnicas de multi-dimensional clustering en base de treinta y ocho variables en más de diez mil muestras de jugadores. Estos roles ayudan en diferenciar jugadores que juegan la misma posición tradicional, y ayudará en producir nuevos comparaciones entre jugadores. Usando diecinueve temporadas de estadísticas de jugadores, modelos fueron desarollados usando tres tecnicas de clustering: Gaussian Mixtures, DBSCAN, y k-Means. Después del desarollo de todos los modelos, uno fue eligido que proporcionó los mejores clusters y fue utilizada para identificar los nuevos roles. Estos nuevos roles se pueden ser utilizadas para identificar reemplazos para ciertos jugadores, contratando a un jugador que realizan el mismo rol, o haciendo comparaciones entre nuevos jugadores en el NBA y los roles historicós que otros jugadores han realizado.

Contents

List of Figures

List of Tables

Dedicated to ....

My family for supporting me as I went on this journey, and always pushing me to be better.

1 Introduction

Contents

1.1 Motivation

In recent years, machine learning and data science have expanded into world of sports, hoping to identify opportunities that teams or coaches have in improving their chances of winning games. Additionally, the realm of sports science is used to model player performances, identifying areas where they excel or need to improve. This concept of modeling player performances was the basis for this thesis, with a focus on finding similarities between players, rather than identifying areas where players perform well or need improvement. By identifying where players are similar, comparisons between new players and old players can be drawn in the hopes of developing better understandings.

1.2 Background

Basketball is one of the world's largest sports, and generally one of the simplest in scope. Two teams of five players each aim to score points by putting a ball into a basket, and at the end of the game the team with the highest number of points wins. This is the game in simplest terms. There exist, however, several different ways to further understand how the game is played, and there exist statistics abound that are collected by the NBA that allow for further investigation.

One area in which investigation is done is into what are they key statistics that allow for the better prediction of game outcomes. [1,2,3,4,5]

[1] Dragan Miljković, Ljubisa Gajić, Aleksandar Kovacevic, and Zora Konjovic. The use of data mining for basketball matches outcomes prediction. pages 309–312, 09 2010. ISBN 978-1-4244-7394-6. DOI: 10.1109/SISY.2010.5647440

[2] Kathleen Jean Shanahan. A model for predicting the probability of a win in basketball. Master's thesis, University of Iowa, 1984

[3] Eftim Zdravevski and Andrea Kulakov. *System for Prediction of the Winner in a Sports Game*, pages 55–63. 01 2010. ISBN 978-3-642-10780-1. DOI: 10.1007/978-3-642-10781-87

[4] Chenjie Cao. Sports data mining technology used in basketball outcome prediction. Master's thesis, Technological University Dublin, 2012

[5] Bernard Loeffelholz, Earl Bednar, and Kenneth Bauer. Predicting nba games using neural networks. *Journal of Quantitative Analysis in Sports*, 5:7–7, 02 2009. DOI: 10.2202/1559-0410.1156

By being able to identify what they key statistics are that are able to predict game outcomes, NBA teams should be able to identify areas in which they must perform better than their opponents. In doing so, teams are able to perform better in the regular season, and hopefully reach the NBA Championship at the end of the year. However, identifying that a team needs to shoot more three-point shots or obtain a higher number of offensive rebounds is not enough. Teams must identify the players that best allow them to achieve those better statistics. They have to be able to not only choose the five players that will start the game, but also the players that come in off of the bench to relieve the starters. A team's coach must be able to identify the abilities of each player, and how they are able to be used to achieve wins.

In general, there are five traditional positions in basketball [6]:

[6] Jr NBA. Basketball positions. https://jr.nba.com/basketball-positions/

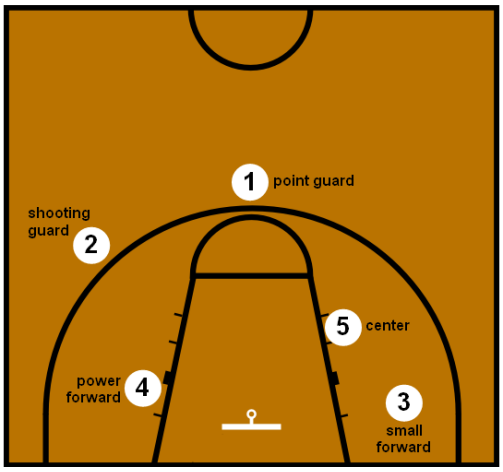


Figure 1.1: Basketball positions

- **Center**: The center is the tallest player on each team, playing near the basket. On offense, the center tries to score on close shots and rebound. But on defense, the center tries to block opponents' shots and rebound their misses.
- **Power Forward**: The power forward does many of the things a center does, playing near the basket while rebounding and defending taller players. But power forwards also take longer shots than centers.
- **Small Forward**: The small forward plays against small and large players. They roam all over on the court. Small forwards can score from long shots and close ones.
- **Shooting Guard**: The shooting guard is usually the team's best

shooter. The shooting guard can make shots from long distance and also is a good dribbler.

- **Point Guard**: The point guard runs the offense and usually is the team's best dribbler and passer. The point guard defends the opponent's point guard and tries to steal the ball.

While these are the simplest way to define a player's position, these definitions cannot describe a player's play style completely, or how they are utilized in a team's composition. Although three players might all play the same position, their abilities on the basketball court can be dramatically different. Being able to differentiate between two players of the same traditional position is key for any basketball coach. This differentiation helps describe the role that a player has beyond their traditional position. We can see this differentiation by looking at the statistics of the players in Table 1.1.[7]

[7] Basketball-reference. https://www.basketball-reference.com/

Name	MP	FG	FGA	3P	3PA	TRB	AST	STL	BLK	PTS
DeAndre Jordan	29.7	4.1	6.5	0.0	0.0	13.1	2.3	0.6	1.1	11.0
Joel Embiid	33.7	9.1	18.7	1.2	4.1	13.6	3.7	0.7	1.9	27.5
Jordan Bell	11.6	1.5	2.8	0.0	0.0	2.7	1.1	0.3	0.8	3.3

Table 1.1: Selected statistics of three centers from the 2018-19 NBA regular season. Definitions can be found in the glossary

While both DeAndre Jordan and Joel Embiid play a large portion of the game's 48 minutes, their offensive output is severely different. Joel Embiid is tasked with shooting almost three times as many shots, while still putting up comparable numbers in terms of total rebounds. A player like Jordan Bell however, who rarely starts, is asked to put in a different performance with his limited minutes. Even with those limited minutes however, he still produces a comparable number of blocks to DeAndre Jordan, showing that his presence as a rim defender is still useful.

From this token example among three centers, it is clear their exist differences even among players that play the same position for their teams. Players are not solely defined by their positions, but also their play styles, and the roles that they fulfill for their teams.

1.3 Objectives

1.3.1 Main Objective

This thesis will aim to show that the positions of players can be further expanded using a multi-dimensional clustering analysis to group similar players together. Instead of the traditional five positions, a model will be developed to identify different classifications of players

that can be used to divide said players based on their play style and re-define the role that they serve on their teams.

1.3.2 *Secondary Objective*

In particular, three different multi-dimensional clustering techniques will be used: DBSCAN, Gaussian Mixture, and K-Means. Models will be developed using player statistics from the 2000-01 NBA season through the 2018-19 NBA season, and comparisons will be done in order to identify the model that provides the best results. Afterwards, after choosing a model, the clusters in the chosen model will be used to identify and define the roles that each clusters' players fill for their teams.

1.4 *Previous Work*

On the topic of clustering, work has been previously done by Patel [8] and Cheng [9]. In both sets of work, the scope of the investigation was limited to only a single season in the case of Patel (2016-17 regular season) and two and a half seasons in the case of Cheng. Furthermore, work by both Cheng and Patel used per-100-possessions statistics, rather than raw statistics. The key difference between the two types of statistics is that the per-100-possessions statistics aim to extrapolate player performances onto a similar baseline, while the raw statistics aim to show a true idea of how a player is used. Two players that share similar per-100-possessions statistics might be viewed the same, but if one player plays 36 minutes per game while the other plays 10, there is a large difference.

These two pieces of work also reduced the dimension of the data set using t-Distributed Stochastic Neighbor Embedding (t-SNE) and Principal Component Analysis (PCA) in the case of Patel, and Linear Discriminant Analysis (LDA) and PCA in the case of Cheng. This facilitates the ability to visualize the different clusters clearly, but information is lost in the reduction of dimension.

In contrast to these two pieces of work focus on per-100-possessions statistics and the reduce the dimension of the resulting models, this thesis uses raw statistics and maintains the dimension of the model. In using raw statistics from 19 seasons of the NBA a more 'pure' model is constructed and is reinforced through the use of more data. This 'pureness' is also produced in maintaining the dimensionality of the data set, which allows for all of the information to be used in the construction of the models.

[8] Riki Patel. Clustering professional basketball players by performance. Master's thesis, University of California, Los Angeles, 2017

[9] Alex Cheng. Using machine learning to find the 8 types of players in the nba. https://medium.com/fastbreak-data/classifying-the-modern-nba-player-with-machine-learning-539da03bb824, 2017

1.5 Thesis Structure

This thesis will start in Chapter 2 with a description on the mathematical theory behind PCA and each of the three clustering techniques. Afterwards follows, in Chapter 3, a description of the data set used and the methods used to pre-process the data. Next, in Chapter 4 the three clustering techniques are compared and the final model is described. Finally in Chapter 5 conclusions of the thesis are reached, and future work is proposed.

2 Mathematical Preliminaries

Contents

This chapter will detail the mathematical theory behind the techniques used through this thesis. First is how the dimension of a model can be reduced using PCA for the display of the models. Afterwards, each of the three modeling techniques and the process behind them is discussed.

2.1 Principal Component Analysis

PCA is one of the more widely used tools in a variety of fields due to its simplicity in extracting relevant information from data sets filled with noise. Although simple in its construction, PCA allows the noisy data sets to be reduced to lower dimensions that can contain more important information about how the structure of the data set.

Suppose their exists a data set represented by $\mathbf{X}$, an $m \times n$ matrix, with m variables, and n samples. In order to find the principal components of $\mathbf{X}$, an orthonormal matrix $\mathbf{P}$ in $\mathbf{Y} = \mathbf{PX}$ must be found such that $\mathbf{\Sigma_Y} = \frac{1}{n}\mathbf{YY}^\intercal$ is a diagonal matrix. In finding such $\mathbf{P}$, the principal components of $\mathbf{X}$ can be found in the rows of said matrix.

The complete details on how this conclusion is reached can be found in work by Shlens [1], however the idea is that by choosing to create $\mathbf{P}$ such that each row $\mathbf{p_i}$ is an eigenvector of $\frac{1}{n}\mathbf{XX}^\intercal$, it follows that $\mathbf{\Sigma_Y}$ must be diagonal. Through $\mathbf{\Sigma_Y}$ it is also shown that the i^{th} diagonal value represents the variance of $\mathbf{X}$ in the i^{th} principal component $\mathbf{p_i}$.

[1] Jonathon Shlens. A tutorial on principal component analysis, 2014

In practice, only the first k principal components are used in the projection of the data set, and therefore $\mathbf{P}$ will be a $k \times m$ matrix, and the projected data set $\mathbf{Y}$ can be found by resolving $\mathbf{Y} = \mathbf{PX}$.

2.2 Density-Based Spatial Clustering of Applications with Noise

DBSCAN is one of the first density-based clustering algorithms, and is useful in discovering clusters where data sets have arbitrary shape or are of large size. This type of algorithm typically performs well when the data set has dense regions separated by regions of low density. DBSCAN in particular builds the clusters based on a connectivity analysis.

DBSCAN is derived from the idea that each cluster must have a minimum cardinality (*MinPts*) within a given radius ϵ. Both of these are hyper-parameters given by the model creator. The neighborhood for a point p is given by,

$$N_\epsilon(p) = \{q \in \mathbf{X} \,|\, dist(p,q) < \epsilon\}.$$

If the point P has cardinality greater than *MinPts*, that is $|N_\epsilon(P)| > MinPts$, then P is considered a core point, and a cluster is developed from that point. This process is repeated until all points in the data set are classified in a cluster or as noise. Those points classified as noise are done so because they are never within a distance ϵ of another point in the data set that was assigned to a cluster.

For more information regarding the process of DBSCAN, or improvements that have been made to the original algorithm refer to work by Khan et al. [2]

[2] Kamran Khan, Saif Ur Rehman, Kamran Aziz, Simon Fong, and S. Sarasvady. Dbscan: Past, present and future. In *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, pages 232–238, 2014. DOI: 10.1109/ICADIWT.2014.6814687

2.3 K-Means Modeling

K-Means is one of the most widely used clustering techniques due to its ease of use and understanding. The general theory is to partition a data set into k partitions. This is done through an iterative process:

1. Create k initial cluster centers (centroids) in the data space
2. Compute the distance from each data point x_i to each cluster center k_j, and assign the point to the closest centroid
3. Recalculate the value for the cluster center by taking the mean of each point assigned to the center k_j
4. Repeat 2. and 3. until there is no more change in assignment

While the theory is simple to understand, there are some pitfalls that can arise using k-Means. Due to the fact that the initial centers are random, each iteration of the k-means model can result in different results, but there do exist methodologies for refining the initial points as shown in work by Bradley and Fayyad [3]. There is also the question of the number of centers to be initialized, but this is easily resolved

[3] Paul S. Bradley and Usama M. Fayyad. Refining initial points for k-means clustering. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, page 91–99, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 1558605568

through hierarchical clustering or through simply trying multiple values for k.

Further restrictions can be put on the k-Means algorithm by constraining certain pairs of data points such that they are always in the same cluster, or such that they are always in separate clusters. [4]

[4] Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schrödl. Constrained k-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 577–584, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558607781

2.4 *Gaussian Mixture Modeling*

Gaussian Mixture clustering, like k-Means aims to partition the data set into a certain number of clusters, or components as they are referred to within the context of Gaussian Mixtures. However, rather than just finding the centers of each of the components, further attributes or features are used to define the components.

The main objective of the Gaussian Mixture is to find an estimate of $\Theta^* = \{\alpha_j^*, \theta_j^*\}_{j=1}^{k^*}$, typically written as $\Theta = \{\alpha_j, \theta_j\}_{j=1}^{k}$, where k is an estimate of the true model order k^*. This usually done by solving the following maximum likelihood (ML):

$$\Theta_{\text{ML}} = \arg\max_{\Theta}\{\log p(\mathbf{X_N} \mid \mathbf{\Theta})\},$$

where N is the number of observations in the data set.

However, all of this is done assuming that the original distribution of the data set is formed from a mixture of k^* Gaussian components:

$$p(\mathbf{x}_t \mid \Theta^*) = \sum_{j=1}^{k^*} \alpha_j^* p(\mathbf{x}_t \mid \boldsymbol{\theta}_j^*)$$

with

$$\sum_{j=1}^{*} \alpha_j^* = 1 \quad \text{and} \quad \forall 1 \leq j \leq k^*,\ \alpha_j^* > 0,$$

where each observation $\mathbf{x_t}$ is a column vector of m-dimensional features. Additionally, $p(\mathbf{x}_t \mid \Theta^*)$ is the j^{th} Gaussian component with $\boldsymbol{\theta}_j^* = \{\boldsymbol{\mu}_j^*, \boldsymbol{\Sigma}_j^*\}$, where $\boldsymbol{\mu}_j^*$ and $\boldsymbol{\Sigma}_j^*$ are the mean vector and covariance matrix of the j^{th} component, respectively. α_j^* is the true mixing coefficient of the j^{th} component.

Like k-Means, often the number of clusters must be pre-determined, however there exist methodologies to determine the model order and features at the same time [5].

[5] Hong Zeng and Yiu-ming Cheung. A new feature selection method for gaussian mixture clustering. *Pattern Recognition*, 42:243–250, 02 2009. DOI: 10.1016/j.patcog.2008.05.030

3 The Data

Contents

3.1 Source of Data

All of the data used in this paper was gathered from BasektballReference.com [1] with use of the Python library Basketball Reference Web Scraper [2]. Further code was developed to acquire data that was not available using the existing library.

[1] Basketball-reference. https://www.basketball-reference.com/

[2] Jae Bradley. Basketball reference web scraper. https://github.com/jaebradley/basketball_reference_web_scraper

3.2 Data Information

In order to build a data set based off of player performance, data for players must be collected. The decision was made to treat a player's performance for separate seasons as unique data points. Furthermore, if a player were to play for more than one team in a season due to a trade or other reason, each individual team that a player played for in a season would be treated as unique.

This was done for two purposes. First, to expand the number of data points available to the model in hopes of getting better performance.

Secondly, each time a player plays for a new team, or in a new season, the role that that player will fill might change. A player in their rookie season will not necessarily be expected to fill the same role as they would as a five-year veteran. Furthermore, a player in the later stages of their career will not have the same output as their younger years. Additionally, a player might be a bench player with limited minutes, but then after being traded might be expected to perform in a different role.

As can be seen in Figure 3.1, over time the number of players participating in a season has increased, and overall there are 10,010 players over the 19 seasons, for an average of around 527 players per individual season.

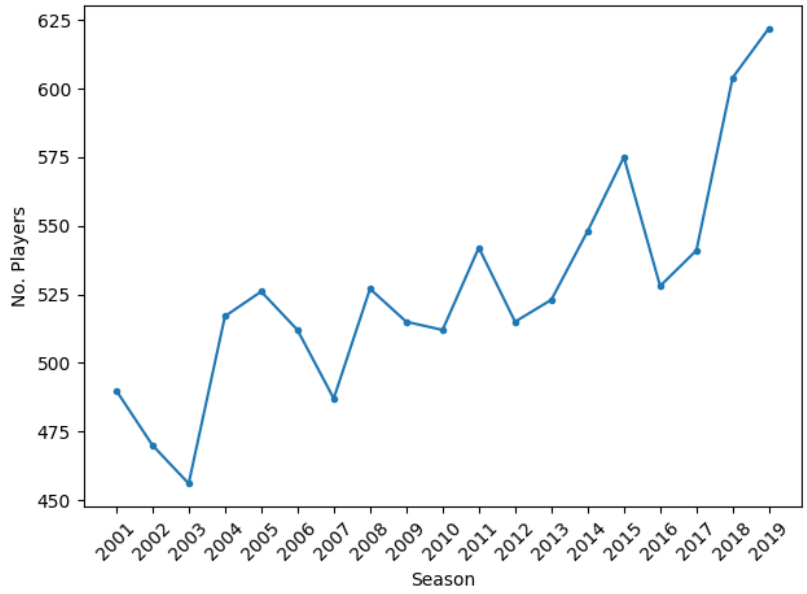


Figure 3.1: Number of players for each season in data set

As for the individual statistics that will be used to construct the model, they come from three separate tables on Basketball-Reference. They are the player's totals, advanced, and shooting statistics. Examples of these tables can be seen in Figure 3.2. Overall there were a total of 56 variables that were imported into the data set for each player.

3.3 *Data Description*

The 56 variables that were imported into the data set can be broken down into five general categories: metadata, calculated statistics, traditional statistics, adjusted statistics, and shooting statistics.

Totals Share & Export ▾ Glossary

Season	Age	Tm	Lg	Pos	G	GS	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	Trp Dbl
2013-14	19	MIL	NBA	SF	77	23	1897	173	418	.414	41	118	.347	132	300	.440	.463	138	202	.683	78	261	339	150	60	61	122	173	525	0
2014-15	20	MIL	NBA	SG	81	71	2541	383	780	.491	7	44	.159	376	736	.511	.496	257	347	.741	100	442	542	207	73	85	173	254	1030	0
2015-16	21	MIL	NBA	PG	80	79	2823	513	1013	.506	28	109	.257	485	904	.537	.520	296	409	.724	113	499	612	345	94	113	208	**258**	1350	5
2016-17 ★	22	MIL	NBA	SF	80	80	2845	656	1259	.521	49	180	.272	607	1079	.563	.541	471	612	.770	142	558	700	434	131	151	234	246	1832	3
2017-18 ★	23	MIL	NBA	PF	75	75	2756	742	1402	.529	43	140	.307	699	1262	.554	.545	487	641	.760	156	597	753	361	109	106	223	231	2014	1
2018-19 ★	24	MIL	NBA	PF	72	72	2358	721	1247	.578	52	203	.256	669	1044	.641	.599	500	686	.729	159	739	898	424	92	110	268	232	1994	5
2019-20 ★	25	MIL	NBA	PF	63	63	1917	**685**	1238	.553	89	293	.304	**596**	945	.631	.589	398	629	.633	140	**716**	856	354	61	66	230	195	1857	4
2020-21 ★	26	MIL	NBA	PF	45	45	1529	468	829	.565	51	169	.302	417	660	.632	.595	307	**447**	.687	80	432	512	277	51	60	169	129	1294	7
Career			**NBA**		**573**	**508**	**18666**	**4341**	**8186**	**.530**	**360**	**1256**	**.287**	**3981**	**6930**	**.574**	**.552**	**2854**	**3973**	**.718**	**968**	**4244**	**5212**	**2552**	**671**	**752**	**1627**	**1718**	**11896**	**25**

(a)

Advanced Share & Export ▾ Glossary

Season	Age	Tm	Lg	Pos	G	MP	PER	TS%	3PAr	FTr	ORB%	DRB%	TRB%	AST%	STL%	BLK%	TOV%	USG%	OWS	DWS	WS	WS/48	OBPM	DBPM	BPM	VORP
2013-14	19	MIL	NBA	SF	77	1897	10.8	.518	.282	.483	4.6	16.3	10.2	12.1	1.7	2.6	19.4	15.0	0.1	1.1	1.2	.031	-2.4	-0.1	-2.5	-0.2
2014-15	20	MIL	NBA	SG	81	2541	14.8	.552	.056	.445	4.5	19.7	12.2	13.1	1.5	2.8	15.6	19.6	2.2	4.0	6.2	.117	-1.2	1.2	0.0	1.2
2015-16	21	MIL	NBA	PG	80	2823	18.8	.566	.108	.404	4.6	20.0	12.4	20.0	1.7	3.4	14.8	22.3	4.2	2.9	7.1	.121	1.3	0.7	2.1	2.9
2016-17 ★	22	MIL	NBA	SF	80	2845	26.1	.599	.143	.486	5.9	22.6	14.3	26.6	2.3	4.7	13.3	28.3	7.9	4.5	12.4	.210	4.9	2.4	7.3	6.7
2017-18 ★	23	MIL	NBA	PF	75	2756	27.3	.598	.100	.457	6.7	25.3	16.0	23.7	2.0	3.3	11.7	31.2	8.3	3.6	11.9	.207	5.3	0.9	6.2	5.7
2018-19 ★	24	MIL	NBA	PF	72	2358	**30.9**	.644	.163	.550	7.3	30.0	19.3	30.3	1.8	3.9	14.8	32.3	8.9	5.5	14.4	**.292**	6.3	**4.1**	10.4	7.4
2019-20 ★	25	MIL	NBA	PF	63	1917	**31.9**	.613	.237	.508	7.7	34.8	22.1	34.2	1.5	3.0	13.2	**37.5**	6.1	**5.0**	11.1	**.279**	7.4	**4.1**	**11.5**	6.6
2020-21 ★	26	MIL	NBA	PF	45	1529	29.0	.631	.204	.539	5.7	28.7	17.6	29.3	1.6	3.5	14.1	32.8	5.2	2.6	7.7	.243	6.0	2.5	8.5	4.1
Career			**NBA**		**573**	**18666**	**23.5**	**.599**	**.153**	**.485**	**5.8**	**24.3**	**15.3**	**23.3**	**1.8**	**3.5**	**14.1**	**27.1**	**42.8**	**29.3**	**72.1**	**.185**	**3.4**	**1.9**	**5.3**	**34.4**

(b)

Shooting Shot location data available for the 1996-97 through 2020-21 seasons. Share & Export ▾ Glossary

									% of FGA by Distance						FG% by Distance						% of FG Ast'd		Dunks		Corner 3s		Heaves	
Season	Age	Tm	Lg	Pos	G	MP	FG%	Dist.	2P	0-3	3-10	10-16	16-3P	3P	2P	0-3	3-10	10-16	16-3P	3P	2P	3P	%FGA	#	%3PA	3P%	Att.	#
2013-14	19	MIL	NBA	SF	77	1897	.414	10.8	.718	.476	.098	.045	.098	.282	.440	.573	.171	.105	.220	.347	.621	.707	.160	61	.127	.067	1	0
2014-15	20	MIL	NBA	SG	81	2541	.491	7.6	.944	.506	.178	.067	.192	.056	.511	.646	.288	.423	.393	.159	.519	.857	.129	95	.159	.143	1	0
2015-16	21	MIL	NBA	PG	80	2823	.506	7.5	.892	.509	.180	.058	.143	.108	.537	.684	.346	.288	.359	.257	.485	.821	.143	141	.321	.343	4	0
2016-17	22	MIL	NBA	SF	80	2845	.521	8.7	.857	.496	.149	.064	.146	.143	.563	.709	.394	.338	.342	.272	.420	.714	.164	194	.111	.200	4	0
2017-18	23	MIL	NBA	PF	75	2756	.529	8.4	.900	.454	.173	.114	.160	.100	.554	.756	.355	.350	.339	.307	.429	.907	.122	161	.107	.333	1	0
2018-19	24	MIL	NBA	PF	72	2358	.578	7.7	.837	.573	.152	.051	.063	.163	.641	.769	.339	.381	.410	.256	.472	.673	.246	279	.044	.222	1	0
2019-20	25	MIL	NBA	PF	63	1917	.553	9.8	.763	.482	.160	.073	.048	.237	.631	.779	.343	.433	.400	.304	.438	.551	.170	197	.034	.500	0	0
2020-21	26	MIL	NBA	PF	45	1529	.565	9.4	.796	.458	.203	.071	.064	.204	.632	.834	.375	.339	.321	.302	.432	.275	.169	133	.047	.375	0	0
Career			**NBA**		**573**	**18666**	**.530**	**8.6**	**.847**	**.496**	**.165**	**.071**	**.114**	**.153**	**.574**	**.733**	**.345**	**.356**	**.355**	**.287**	**.458**	**.639**	**.165**	**1261**	**.095**	**.277**	**12**	**0**

(c)

Figure 3.2: Data tables for Giannis Antetokounmpo. (a) shows the totals table, (b) shows the advanced statistics table, (c) shows the shooting table

3.3.1 *Metadata and Calculated Statistics*

The variables in these two categories contain information relevant to identifying the player (metadata) or metrics that were devised by other parties (calculated statistics). In the end these variables were discarded when building the clustering models due to being literal descriptors in the case of the metadata or too abstract in the case of the calculated statistics.

3.3.2 *Traditional statistics*

The variables in this category are those traditionally found in a standard box score for a game. These include shooting statistics like attempted field goals, attempted three point field goals, attempted free throws, etc., along with other statistics such as points, total rebounds, steals, blocks, turnovers, etc.

The key thing to note about these statistics is that they are all in the raw totals for an entire season. Rather than use a per-36-minute or a per-100-possession extrapolation, the raw totals were used as a way to

give a true estimate as to what the player actually provided to the team in terms of raw output. The only thing to note is that the stats were averaged out on a per-game basis as will be discussed in Section 3.4.4.

A note about these statistics, there exists a large amount of linear dependence in the data set. For example, in the data set there exist three rebounding traditional statistics: offensive rebounds, defensive rebounds, and total rebounds. Due to the simple nature of offensive and defensive rebounds combining into total rebounds, it will be possible to eliminate some variables.

With all of these being raw sums, the values for these variables are all integers greater than or equal to 0, with no explicit upper bound.

3.3.3 *Adjusted Statistics*

In order to increase the interpretability of some of the traditional statistics, some calculations were made, leading to the adjusted statistics. Rather than just looking at the raw rebounding statistics for a player, it might be interesting to look at how often the player is getting the rebounds available to them. This leads to the calculation of the statistics offensive rebound percentage, defensive rebound percentage, and total rebound percentage. This same process is used for shooting leading to true shooting percentage, steals and steal percentage, etc.

While there still exists linear dependence between these statistics, there is enough of a difference in how they are calculated, due to several using opponents' statistics in their calculation, that they were not thrown out.

Each of these adjusted statistics is seen as a percentage or rate, and as such lie in a range from 0 to 1.

3.3.4 *Shooting Statistics*

The shooting statistics are interesting because they allow for the ability to categorize players based off of where they are choosing to take their shots, and how often they perform in those areas. There are three different types of shooting statistics that are included.

The first is shot selection, from what distance a player is choosing to take their shots. These are broken up into the 0-3ft, 3-10ft, 10-16ft, 16-xxft, and three-point ranges. The xxft represents the fact that the three point arc on the basketball court is not equidistant from the hoop, so the 16-xxft range is all two-point shots from a distance greater than 16 ft.

The second is shot making ability, how often a player makes a shot taken in the previously mentioned ranges. This is where the data set gets interesting with missing values. Due to the nature of these statistics, if a player never takes a shot from three-point range, the

percentage of shots made in that range is infinity. This comes from dividing the total number of shots made in that range (0) by the total number of shots attempted in that range (0), giving infinity due to being unable to divide by 0. However, the data set just leaves these entries as blank.

Finally are the miscellaneous shooting statistics that include how often a player is being assisted on their baskets, and how often they are dunking the basketball. Again, there are missing data points here occurring when a player never takes a three-point shot, and therefore are never assisted on their three-point shots.

All of these statistics are rates or percentages, and as such lie in a range from 0 to 1. For each player, the sum of shot selection variables will sum to 1, while the shooting percentages will vary. As a whole though, the average shooting percentages drop as the distance increases.

Furthermore, there are a total of seven variables that have missing data in this section. They are the five shooting percentage variables, along with the two variables that measure how often a shot made by a player is assisted.

3.4 *Pre-Processing the Data*

In order to use the data in the models, some pre-processing must be done in order to have the models behave better.

3.4.1 *Missing Values*

The first question is what to do with the missing values that occur in the shooting percentage variables. There are three options that were considered: drop the variables that have missing data, set the missing data equal to the mean value of the variable, or to set the missing data equal to 0.

The first option was not implemented due to the fact that information would be lost by removing the variables in question. When considering the fact that the seven variables have interesting information surrounding a player's shooting style, it was best to find a different method.

The second option of setting the value equal to the mean did not make any logical sense. If a player does not have any shots from a certain range, there are two possible explanations. Either the player just overall does not have a large enough volume of shots or the player knows that they are not skilled enough to make the shots consistently. The first of these explanations is handled by Section 3.4.2. However taking the second explanation into account, it does not make sense to reward a player who chooses not to take shots from a specific region by

boosting their percentages to the mean.

As such, the decision on how to handle the missing data fell to the third option, setting the missing data equal to the value 0. This works logically in the sense that a player who has not taken any shots from a certain range is actually shooting 0% from that range. It also works functionally because it allows for the models to run smoothly without having to do any fancy imputations of the missing data.

3.4.2 *Trimming Data*

With the regular season of the NBA consisting of 82 games, the question arises how many games is truly necessary in order to determine how a player's role is defined. With this question in mind, four different values were selected to try: 5, 10, 15, and 20. The number of players with at least the minimum requirement can be seen below.

Minimum Number of Games	Number of Players
0	10010
5	9410
10	8826
15	8337
20	7871

Table 3.1: Number of players satisfying the minimum number of games played

3.4.3 *Dropping Variables*

While the metadata and calculated statistics as described in Section 3.3.1 have already been dropped, there are still a couple of variables that do not really provide much information, or are superfluous.

The first of these are the games played and games started statistics. While these are interesting in showing a player's role on a team in a season-wide view, they do not show any information on a game-to-game view. However, the games played statistic will be useful in helping with the standardizing of the raw traditional statistics.

Secondly, there is a duplicate variable in the data set. The variable three point attempt rate, which shows up in the adjusted statistics is the exact same value as variable in the shooting statistics which describes how many shots a player makes from three-point range. Due to this fact, only one of them is needed, and the variable three point attempt rate was dropped.

Finally, due to the fact that total rebounds is simply the sum of offensive and defensive rebounds, the variable total rebounds was dropped.

After these variables have been dropped, 38 variables in total exist for the use in modeling.

3.4.4 Standardizing and Normalizing Variables

Due to the fact that a large amount of the variables is dealing with raw sums, decisions were made on how to handle said variables, along with how to treat all variables equally.

The first order of business was to take all of the traditional statistics as referenced in Section 3.3.2 and standardize them to a game scale. In order to do this each of the variables was divided by the number of games played in the season, leaving the data with averages for the whole season for each of our variables.

However, in order to combat the discrepancy in ranges for the variables, and to treat each variable equally, a normalization process had to be done. The chosen normalization was a min-max method with the minimum value being set to 0 and the maximum being set to 1.

3.4.5 Dimensionality

When pre-processing the data, the question arose whether or not the reduction of the dimension of the data set using PCA would be an agreeable option. Tests were made and the ultimate conclusion was decided not to use PCA to reduce the dimension in the creation of the models, only for the visualization.

This was done due to the fact that with only two dimensions, the explained variance by the two components was relatively low. The hopes were that the explained variance with two components could be at least 80%, but this was not the case. The actual explained variances for the four minimum threshold of games is as follows:

Min No. Games	Explained Var. Comp. 1	Explained Var. Comp. 2	Total Explained Var.
5	.3062	.2185	.5247
10	.3095	.2275	.5370
15	.3221	.2317	.5538
20	.3213	.2346	.5560

Table 3.2: Explained variance of first two components of PCA reduction on data sets

4 The Problem

Contents

In this chapter, the prospective modeling techniques will be introduced, followed by the resulting output of each technique. Afterwards, a further expansion on the results of the selected model will be discussed.

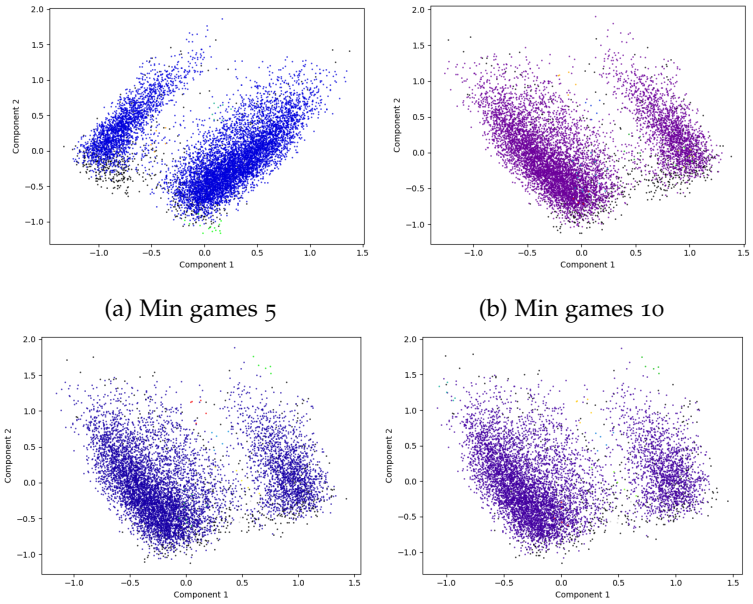
4.1 Prospective Models

In order to determine the best methodology on which to build the clustering model, three different algorithms were used. The first DBSCAN, a density based algorithm where the number of clusters outputted is determinant of some hyper-parameters chosen. Next is a Gaussian Mixture model, which aims to partition the data set into a determined number of clusters, but the assignment of each point in the data set is probabilistic, and each cluster has the shape of a Gaussian distribution. Finally is k-Means which like the Gaussian Mixture model partitions the data into k clusters in which each point in the data set belongs to the cluster with the nearest centroid, defined as the mean point of that cluster.

4.2 Density-Based Spatial Clustering of Applications with Noise

There are two main hyper-parameters that are used in building a DBSCAN model, the value epsilon which determines the radius around which to search for other observations, and the minimum number of samples in a neighborhood that must exist for a new cluster to be

formed. For the purposes of this thesis, values of epsilon were tried from 0.45 to 0.65, along with a minimum number of samples of either 2, 3, 4, or 5.



(a) Min games 5

(b) Min games 10

(c) Min games 15

(d) Min games 20

Figure 4.1: DBSCAN models of different minimum games played, epsilon = 0.55, min samples = 5

In the above figure, the general behavior of the change in minimum number of games is shown. As the minimum number of games played increases, the behavior of the data set becomes less clustered, as can be seen in Figure 4.2. As the minimum number of games played increases from left to right, the silhouette score generally decreases, showing that the clusters are becoming less well defined.

By counting the number of players that belong to each cluster in one of the models shown in Figure 4.1, a problem becomes apparent with DBSCAN when applied to this data set. Due to the nature of the way that DBSCAN builds the clusters, if the data set is relatively condensed, large clusters will form, and only those points which are far enough away from the main cluster, but still close to other observations will be put into separate clusters. What DBSCAN does perform well in is identifying small clusters of players that perform similarly.

For instance, the seven players in Cluster 2 as detailed by Table 4.1 contain 5 seasons of Ben Wallace (2001, 2003, 2004, 2005, and 2007 seasons) along with 2 seasons of Marcus Camby (2008 and 2010 seasons). While it is interesting to identify similar players in such a manner, for the purposes of identifying overarching roles of players, DBSCAN does not serve.

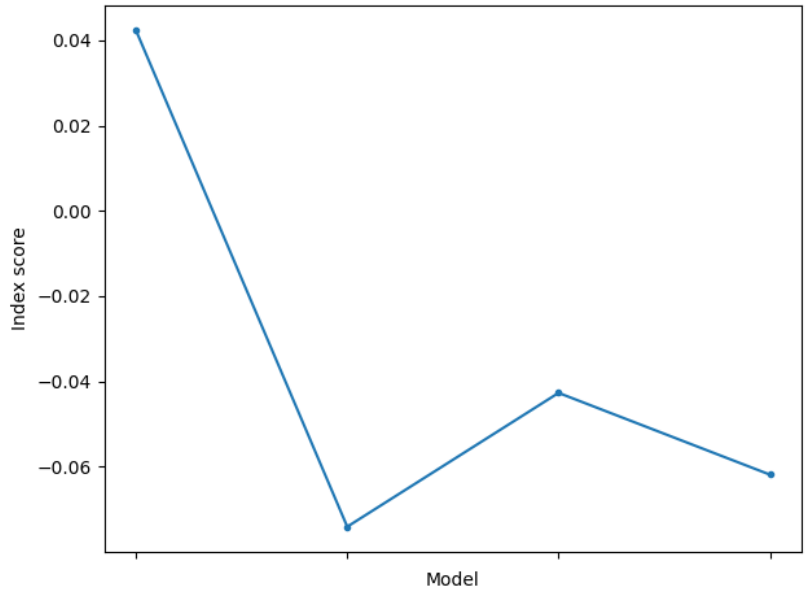


Figure 4.2: Silhoutte index score for DBSCAN models of different minimum games played, epsilon = 0.55, min samples = 5

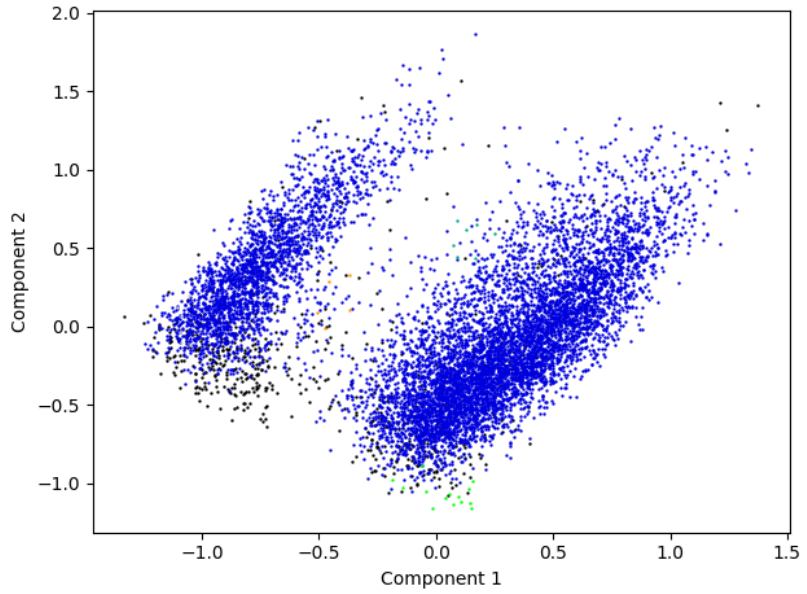


Figure 4.3: DBSCAN model, min games = 5, epsilon = 0.55, min samples = 5

4.3 *Gaussian Mixture*

The Gaussian Mixture allows for the partitioning into a specific number of clusters based off a hyper-parameter. In this thesis the number of clusters tried ranged from 3 through 15. Also, the models

Cluster	No. Players
Noise	648
Cluster 1	8730
Cluster 2	7
Cluster 3	14
Cluster 4	5
Cluster 5	6

Table 4.1: Number of players belonging to each cluster in Figure 4.3

were developed as a 'full' mixture, wherein each cluster has its own covariance matrix defining the distribution.

By looking at Figure 4.4 it is clear that there are some separate clusters defined, and by looking at Table 4.2 the clusters seem to make sense. Those in Cluster 3 seem to be the main shooters and ball distributors as indicated by the large number of field goal attempts (FGA) and assists (AST). Clusters 5 and 9 appear to be defensive specialists that get a large amount of rebounds (ORB and DRB) along with blocks (BLK), while those in Cluster 9 have more ability to shoot from three-point range (3pFGA) when compared to those in Cluster 5. Then there are players like those in Cluster 2 who play limited minutes (MP) while still being productive on the defensive end with their rebounds.

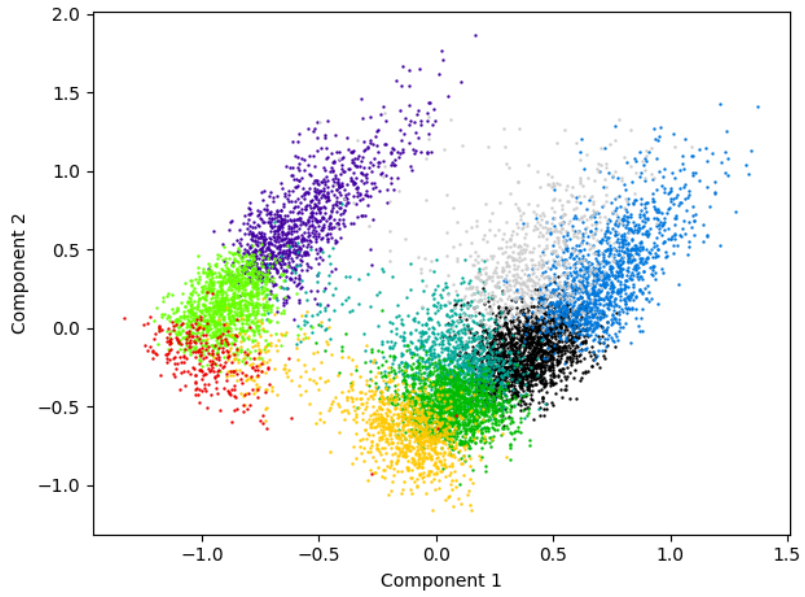


Figure 4.4: Gaussian mixture model with 9 clusters

The main issue with the Guassian Mixture models though is the interpretability for a lay-person who would want to see what role a new player would fit into. This is due to the fact that the calculations for the

Cluster	MP	FGA	3pFGA	ORB	DRB	AST	BLK	PTS
Cluster 1	12.12	3.82	1.67	0.44	1.37	1.18	0.18	4.39
Cluster 2	9.62	2.42	0.02	1.07	1.74	0.42	0.47	3.13
Cluster 3	33.48	15.34	4.18	0.99	3.79	4.77	0.41	19.29
Cluster 4	26.94	9.66	3.71	0.68	2.80	3.20	0.30	11.70
Cluster 5	22.67	7.08	0.03	2.32	4.49	1.18	1.03	9.32
Cluster 6	19.63	6.50	2.61	0.61	2.13	2.01	0.25	7.69
Cluster 7	16.55	4.79	0.36	1.26	2.47	1.31	0.50	5.93
Cluster 8	5.97	1.94	0.73	0.31	0.70	0.68	0.12	2.19
Cluster 9	29.20	11.49	1.09	2.10	5.25	2.49	1.01	14.59

Table 4.2: Selected statistics for means of Gaussian Mixture model with 9 clusters

probability of belonging to each cluster is not easily understood. The idea of clustering players together based on their role is so that anybody can understand where a player fits in, but if a player is actually 30% likely of being role 2 and 35% likely of being role 5, the line between the roles is less clear.

4.4 *K-Means*

Thankfully, with k-means it is clearly defined what the boundaries are for each cluster, and there is no confusion of having players being probabilistically assigned a role. Just like the Gaussian Mixture, there is only one real hyper-parameter, the number of clusters chosen. In this thesis, like with the Gaussian Mixture, the number of clusters tried ranged from 3 through 15.

Similar to DBSCAN, the silhouette score of the models can be used as an indicator of how well defined the clusters in the model are. In Figure 4.5 each consecutive four points represents the increase in minimum number of games from 5 to 20, and each set of four represents an increase in the number of clusters. While the general trend is that the number of clusters decreases how well defined the clusters are, in general increasing the minimum number of games to 15 or 20 helps in having the clusters better defined.

With this in mind in order to try and identify the best number of clusters for a k-means model, a graph of the inertia of each model can be developed like Figure 4.6. While there is easily identifiable elbow that would indicate a clear candidate for the number of clusters, 9 clusters seems to be a point where the decrease in inertia slows.

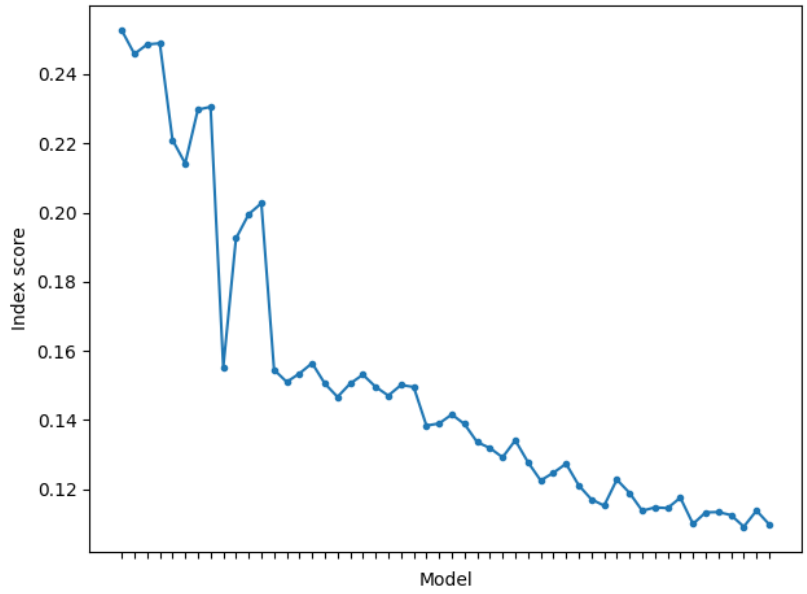


Figure 4.5: Silhouette index scores of k-means models

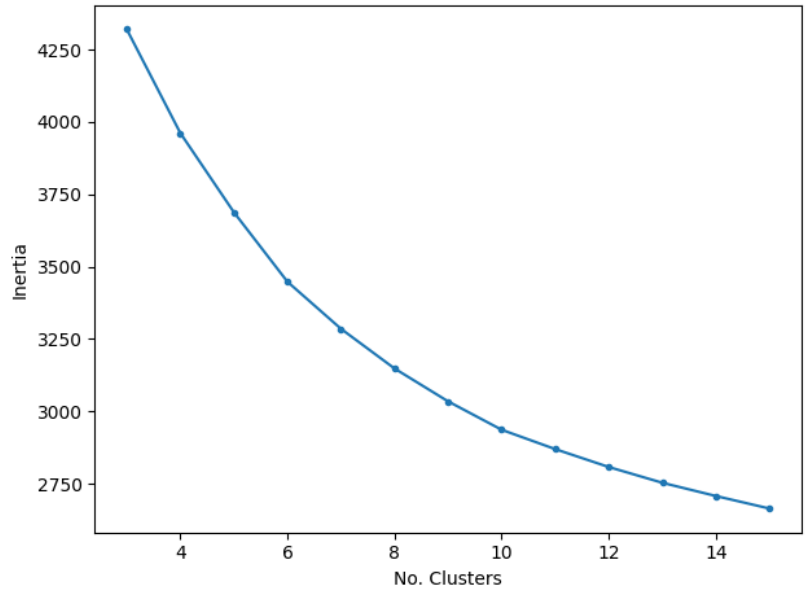


Figure 4.6: Inertia of k-means models with min games = 20

4.5 *The Model*

Ultimately, the chosen model is the k-means model depicted in Figure 4.7. This model provides the best all-around model of player roles by having clearly defined boundaries, while also using a smaller

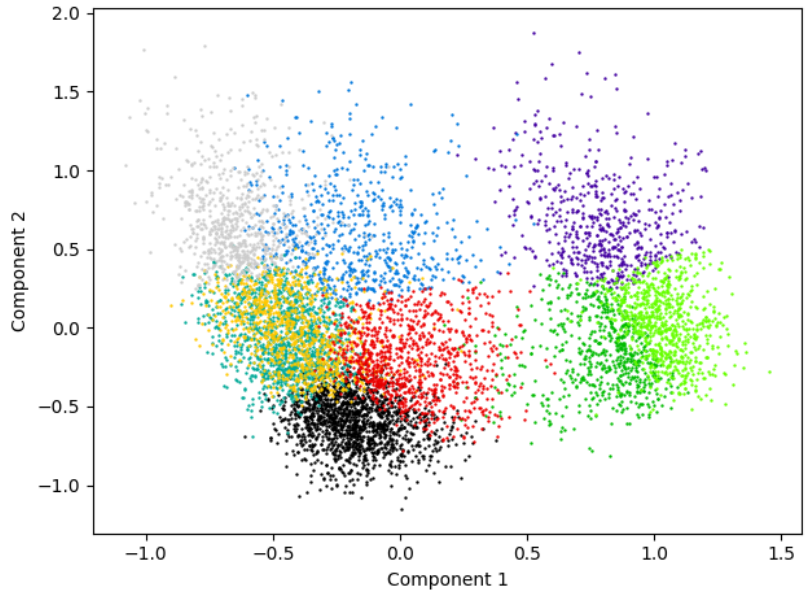


Figure 4.7: k-Means model with 9 clusters, min games = 20

player pool in which players have had more games to have their roles defined on their teams. The number of players in each of the clusters can be seen in Table 4.3.

Cluster	No. Players
Cluster 1	1502
Cluster 2	559
Cluster 3	652
Cluster 4	1200
Cluster 5	621
Cluster 6	775
Cluster 7	942
Cluster 8	924
Cluster 9	696

Table 4.3: Number of players belonging to each cluster in Figure 4.7

Those in cluster 1 can be defined as players who play low minutes but shoot a large number of their shots from three-point range. They also are less likely to get offensive rebounds due to playing on the wing where their shots come from. A noticeable player would be Pat McCaw who played for the Golden State Warriors in 2017 and 2018, and the Toronto Raptors in 2019.

Those in cluster 2 are noticeable for their rebounding, with a large amount of both offensive and defensive rebounds, showing that they like to play inside more often than not. This is also exhibited by their distinct lack of shots taken from three-point range. Ben Simmonds from

the Philadelphia 76ers in 2018 and 2019 is a player in this role.

Those in cluster 3 are high usage players, shooting a lot and distributing the ball relatively well. They also rebound well, showing their versatility around the court. The players in this role include Joel Embiid of the 76ers, Nikola Jokic of the Denver Nuggets, and Jusuf Nurkic of the Portland Trail Blazers in the 2019 season.

Those in cluster 4 are three-point shooters who play a large percentage of the game. They are overall well rounded, although not an overbearing presence on the defensive end. Austin Rivers playing for the Washington Wizards and Houston Rockets in 2019, along with Kyle Kuzma playing for the Los Angeles Lakers in 2018 and 2019 fit this role.

Table 4.4: Selected statistics for centroids of k-means model with 9 clusters, min games = 20

Cluster	MP	FGA	ORB	DRB	AST	BLK	PTS	Usage %	% FGA 0-3ft	% FGA 3pt
Cluster 1	11.22	3.59	0.34	1.14	1.09	0.14	4.07	0.13	0.21	0.40
Cluster 2	26.51	8.75	2.68	5.45	1.44	1.21	11.52	0.16	0.47	0.00
Cluster 3	30.21	12.44	2.13	5.63	2.37	1.00	15.87	0.19	0.34	0.12
Cluster 4	25.56	8.92	0.72	2.81	1.93	0.34	10.85	0.15	0.20	0.45
Cluster 5	11.16	3.29	0.91	1.68	0.72	0.36	3.81	0.13	0.34	0.02
Cluster 6	12.72	3.02	1.50	2.44	0.52	0.69	4.23	0.11	0.61	0.01
Cluster 7	23.11	8.11	0.49	1.99	4.07	0.17	9.49	0.16	0.24	0.29
Cluster 8	16.52	5.08	1.19	2.53	1.06	0.48	6.21	0.13	0.39	0.16
Cluster 9	33.41	15.58	0.88	3.58	5.53	0.37	19.84	0.23	0.26	0.27

The role of those players in cluster 5 is a player who fills in where needed, although not overly flashy. They will often choose to take mid-range shots rather than those in three-point range, and will not have a large output on the scoring end. Shaun Livingston of the Warriors in 2019 and 2018 exemplifies this role.

Cluster 6 is defined by the big man, with limited minutes who is a presence around the rim. Even with limited minutes they put up respectable numbers for rebounding, and more often than not will take their shots from near the hoop. The Plumlee brothers Mason, Marshall, and Miles all perform well in this role.

Players in cluster 7 are known for their ball distribution. Although they might not always put up spectacular numbers, they are able to facilitate their teammates by providing assists. Rajon Rondo over several years has been known for his performances as this role.

Another role that is not necessarily flashy, but yet still does work are those in cluster 8. They work for their team in provide in ways do not always show up in the box score. Andre Iguodala for the Warriors in 2018 and 2019 along with Giannis Antentokounmpo in his rookie season in 2014 performed in this role.

Finally, in cluster 9 are the leaders of the team. They command most of the ball, and generally perform. They also provide for their teammates through their distribution. Players like LeBron James, Kyrie Irving, James Harden, Kevin Durant, and Steph Curry are all examples of players who fulfill this key role for their teams.

5 Conclusions and Future Work

Contents

5.1 Conclusions

Through this thesis, a model that defines nine new roles of players in the NBA was developed based off of 38 player statistics from 19 seasons spanning 2001 through 2019. These roles aim to help broaden the scope of a player's position past the traditional five positions, and to fully encapsulate what a player provides for their team.

Within these roles is the 'leader', where players like Steph Curry, James Harden, and Lebron James command the ball, provide a large amount of offensive output, while also facilitating their teammates through their ball distribution. There is also the role of the 'high usage big man' which describes players like Joel Embiid and Nikola Jokic. While also very important for their teams, this role is separate from the 'leader' because the players who fulfill this role tend to perform better in obtaining rebounds. Another role is the 'three-point specialist' in which players like Austin Rivers and Kyle Kuzma take almost 50% of their shots from three-point range. These players have one of the main purposes of spacing out the court for the rest of their teammates and creating space.

With these nine new roles quantitatively defined through the model produced in this thesis, comparisons between players no long has to rely solely on the 'eye-test' or by using the traditional positions. Now, players can be compared based on their play styles, and similarities between new players that enter the NBA and historical performances can be drawn. Furthermore, the progression of a player and the roles that they have filled over the course of their career can be seen. A player in their rookie season is not likely to fulfill the role of a 'leader', but as they progress in their ability it is possible that their role changes to that

of ‘leader’.

5.2 Future Work

Ideally these roles could be further developed into identifying the best composition of a team, and how to allocate minutes. Being able to identify the importance of each role in an overall team composition and winning could lead to general managers better replacing or acquiring players that would fill holes in their team.

Further work could also be done in expanding the number of variables available, or experimenting in reducing the number of variables due to linear dependence. Although each variable itself tells a story of a player’s performance, it might be superfluous when the combination of two other variables provide the same information. Feature selection is a constant problem in data science, and being able to successfully prune the data set down and still provide clear and concise information is key.

Bibliography

Basketball-reference. https://www.basketball-reference.com/.

Jae Bradley. Basketball reference web scraper. https://github.com/jaebradley/basketball_reference_web_scraper.

Paul S. Bradley and Usama M. Fayyad. Refining initial points for k-means clustering. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, page 91–99, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 1558605568.

Chenjie Cao. Sports data mining technology used in basketball outcome prediction. Master's thesis, Technological University Dublin, 2012.

Alex Cheng. Using machine learning to find the 8 types of players in the nba. https://medium.com/fastbreak-data/classifying-the-modern-nba-player-with-machine-learning-539da03bb824, 2017.

Kamran Khan, Saif Ur Rehman, Kamran Aziz, Simon Fong, and S. Sarasvady. Dbscan: Past, present and future. In *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, pages 232–238, 2014. DOI: 10.1109/ICADIWT.2014.6814687.

Bernard Loeffelholz, Earl Bednar, and Kenneth Bauer. Predicting nba games using neural networks. *Journal of Quantitative Analysis in Sports*, 5:7–7, 02 2009. DOI: 10.2202/1559-0410.1156.

Dragan Miljković, Ljubisa Gajić, Aleksandar Kovacevic, and Zora Konjovic. The use of data mining for basketball matches outcomes prediction. pages 309–312, 09 2010. ISBN 978-1-4244-7394-6. DOI: 10.1109/SISY.2010.5647440.

Jr NBA. Basketball positions. https://jr.nba.com/basketball-positions/.

Riki Patel. Clustering professional basketball players by performance. Master's thesis, University of California, Los Angeles, 2017.

Kathleen Jean Shanahan. A model for predicting the probability of a win in basketball. Master's thesis, University of Iowa, 1984.

Jonathon Shlens. A tutorial on principal component analysis, 2014.

Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schrödl. Constrained k-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 577–584, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558607781.

Eftim Zdravevski and Andrea Kulakov. *System for Prediction of the Winner in a Sports Game*, pages 55–63. 01 2010. ISBN 978-3-642-10780-1. DOI: 10.1007/978-3-642-10781-87.

Hong Zeng and Yiu-ming Cheung. A new feature selection method for gaussian mixture clustering. *Pattern Recognition*, 42:243–250, 02 2009. DOI: 10.1016/j.patcog.2008.05.030.

Glossary

% FGA 0-3ft Percent of field goals attempted within 3 feet.

% FGA 3pt Percent of field goals attempted from three-point range.

3P Three-point field goals made.

3PA Three-point field goals attempted.

3pFGA Number of field goals attempted from three-point range.

AST Assists.

BLK Blocks.

DBSCAN Density-Based Spatial Clustering of Applications with Noise.

DRB Defensive Rebounds.

FG Field goals made.

FGA Field goals attempted.

LDA Linear Discriminant Analysis.

MP Minutes played.

NBA National Basketball Association.

ORB Offensive Rebounds.

PCA Principal Component Analysis.

PTS Points.

STL Steals.

t-SNE t-Distributed Stochastic Neighbor Embedding.

TRB Total rebounds.

Usage % Usage percentage.